Data centre for scientists and institutes in the Netherlands: needs and opportunities

Jeroen Rombouts, Alenka Princic, and Egbert Gramsbergen

TU Delft, Netherlands (j.p.rombouts@tudelft.nl)

Upon an initiative of the libraries of the three Dutch Technical Universities, a research data center is being built for science- and technology research data in Netherlands, called 3TU.Datacentrum (http://datacentrum.3tu.nl). The solutions for different needs and usage of data curation in geosciences and other disciplines are demonstrated by two cases.

The first case is a complex collection of simple datasets consisting of hydrological measurements in a river basin in Luxemburg with several project partners collecting data with several instruments, on several locations, over a long period of time (starting in 2003) and continuing to this day. Due to multiple data users there was a need for central storage to share the data, and for normalization to enable easy combination of data sets. Furthermore the researchers involved believe that collected data should be 'published' as they are valuable to researchers outside of their community in multidisciplinary disciplines such as climate research. Above all, the measurements are not repeatable, which justifies investments in the preservation of the data.

The second case is a simple collection of complex datasets consisting of measurements of the International Research Centre for Telecommunications and Radar (IRCTR). Their researchers have developed meteorological radar called IDRA (IRCTR Drizzle Radar) aimed at obtaining high resolution temporal and spatial measurements of all sorts of precipitation including those non-detectable by standard meteorological radar, like drizzle. In contrast to the first case, these data are from a single instrument and from a single organization but the size of the data collection made sharing with other scientists and reliable backups difficult.

3TU.Datacentrum offered solutions for both cases. Solutions were as standardized as possible yet different where required. Common solutions to both collections are:

• The data model. Datasets, instruments, locations and even periods of time are all objects in their own right, with their own metadata and interconnected through rdf relationships. These relations are exploited in the user interface for navigation and inclusion, e.g. of instrument and location metadata in the display of datasets. They are also used to generate ORE Resource Maps.

• The file format NetCDF. This is an open standard, primarily for sets of multidimensional arrays with metadata included, enabling interoperability and easy use. The NetCDF format can be used by common tools and programming libraries.

Differences in solutions to both collections are in the functionality offered to the user. Darelux is stored in a Fedora repository where datasets of any format are allowed. Extra functionality arises from the fact that alongside the NetCDF an xml version (ncml) is stored that can be converted to other formats like Excel.

The IDRA data are stored on a server hosted by 3TU.Datacentrum supporting the OPeNDAP protocol, which means that internal metadata of NetCDF files are visible and queries within, and in some cases across, datasets are possible. The IDRA dataset contains several raw data files (time samples), one or several processed data files for different radar ranges and a quick look file containing a summary of the rain conditions, operating time and range setting of the radar. This information is also available as a user-friendly diagram which is displayed alongside the other metadata.

Both Darelux and IDRA data collections and descriptive information are now safely stored and available to the scientific community for future research.